Modelling count data extending Poisson regression

Erick A. Chacón-Montalván[1]

[1]*Department of Statistics, National University of Engineering, Peru*

Abstract

Place your abstract here. Remember to include an introduction, description of the problem, goal of the article, description of the methodology, final statement.

Keywords: Mixed models, Poisson regression, random effects, zero-inflation.

About the project

In this project, we work with Poisson models to analyse count data. The Poisson model is a particular case of generalized linear models (Booth and Hobert, 1999; Breslow and Clayton, 1993; McCulloch, 1997). In particular, we will deal with two topics of interest (i) over-dispersion and (ii) zero-inflation. The former occurs when we observed data with empirical variance much bigger the theoretical variance; while the later occurs when the chance of occurrence of zero values is much higher than the expected.

For the first topic, we will consider the modelling of cases of COVID-19 in different districts of Lima (`01-lima-over-dispersion.csv`); while for the second application, we will consider the modelling of a neglected disease such as dengue fever (`02-lima-zero-inflation.csv`). Select one topic of interest and perform the corresponding analysis. Describe the problem, present the methodology and discuss the results in a small article. The data can be found at `https://gitlab.com/ErickChacon/course-computational-statistics-julia`.

Poisson model

Let consider Y_i the number of cases of a disease in district $i = 1, \ldots, n$. A Poisson regression model is usually defined as

$$\begin{aligned} Y_i &\sim \text{Poisson}(\lambda_i = N_i R_i), \\ \log(R_i) &= \boldsymbol{x}_i' \boldsymbol{\beta}. \end{aligned}$$

Where $E(Y_i) = \lambda_i$ is equals to the standardized factor N_i times the risk R_i. $\boldsymbol{\beta}$ are the regression coefficients, of the predictors $\boldsymbol{x}_i$, used to determine the risk R_i. Given a set of observations $y_1, \ldots, y_n$, inference can be done by obtained the log-likelihood function $l(\boldsymbol{\beta})$ and maximizing it.

Over-dispersion

An important characteristic of a Poisson random variable Y is that the mean and variance are the same $E[Y] = V[Y] = \lambda$. In many cases, we might observe count data where this property does not hold, and we observe that $V[Y] >> \lambda$ (over-dispersion). In such cases, we can introduce a latent random variable (or random effects) to be able to fit over-disperse data

$$\begin{aligned} Y \mid Z &\sim \text{Poisson}(\lambda), \\ \log(\lambda) &= Z, \\ Z &\sim \text{Normal}(0, \sigma_z^2). \end{aligned}$$

The resulting random variable Y is more flexible and does not require $E[Y] = V[Y]$.

Using this hierarchical definition, we can easily extend the Poisson model to take into account the over-dispersion as follows:

$$\begin{aligned} Y_i &\sim \text{Poisson}(\lambda_i = N_i R_i), \\ \log(R_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} + Z_i, \\ Z_i &\sim \text{Normal}(0, \sigma_z^2). \end{aligned}$$

Notice that inference of this model for a given set of observations $y_1, \ldots, y_n$ is not as simple as the previous model because the likelihood function $L(\boldsymbol{\beta}, \sigma_z^2)$ is not tractable. However, inference can be still done using Monte-Carlo approximation, stochastic optimization, expectation-maximization algorithm, Markov chain Monte Carlo, and others.

Zero-inflation

In some applications, we observe several zero values which are not modelled adequate with the Poisson probability density function. In those cases we can use a zero inflated Poisson (ZIP) distribution such as

$$Y = \begin{cases} 0, & 1-\pi \\ \text{Poisson}(\lambda), & \pi. \end{cases}$$

The resulting density function of Y is

$$\Pr(y) = (1-\pi)1_{y=0} + \pi \exp(-\lambda)\lambda^y / y! \quad \text{for } y = 0, 1, \ldots.$$

The ZIP random variable can also be defined using a latent random variable $Z \sim \text{Bernoulli}(\pi)$ such as

$$Y \mid Z = \begin{cases} 0, & Z = 0 \\ \text{Poisson}(\lambda), & Z = 1. \end{cases}$$

Using the ZIP random variable, we can extend the Poisson regression model to

$$\begin{aligned} Y_i &= \begin{cases} 0, & 1-\pi_i \\ \text{Poisson}(\lambda_i = n_i R_i), & \pi_i. \end{cases} \\ \log(R_i) &= \boldsymbol{x}_i'\boldsymbol{\beta}, \\ \text{logit}(\pi_i) &= \boldsymbol{w}_i'\boldsymbol{b}. \end{aligned}$$

Where π_i is the probability of coming from a Poisson distribution, which are modelled using the predictors $\boldsymbol{x}_i$ with regression coefficients $\boldsymbol{b}$. Given a set of observations $y_1, \ldots, y_n$, inference can be done by obtained the log-likelihood function $l(\boldsymbol{\beta}, \boldsymbol{b})$ and maximizing it; however it can be simplified by using Monte-Carlo approximation, stochastic optimization, expectation-maximization algorithm, Markov chain Monte Carlo, and others.

1. Introduction

- Present the importance of modelling count data, specially for modelling disease cases.
- Introduce the use of Poisson regression and why to use it.
- Explain the limitation of Poisson regression with respect to over-dispersion and zero zero-inflation.
- Present the goal of the article: extend classical Poisson models to fit more adequately count data.
- Provide the ideas of how you will achieve these goals and the applications you will use
- In one paragraph describe the following structure of the article.

2. Methods

2.1. Poisson regression models

- Present the Poisson regression model. Explain each term and additional properties.
- Provide the likelihood function and explain how inference can be done.
- Describe the limitations of the classical Poisson regression model.
- How would you obtain reliable confidence intervals?

2.2. Poisson regression model with random effects

- Present the Poisson model with random effects. Explain each term and additional properties.
- Explain why this model is able to model over-dispersion. You can use simulations to explain it.
- Explain the way you will make inference for this model.
- How would you obtain reliable confidence intervals?
- Provide a test to check over-dispersion, if necessary use Monte Carlo tests.

2.3. Zero-inflated Poisson regression model

- Present the zero-inflated Poisson model. Explain each term and additional properties.
- Explain why this model is able to model zero-inflation. You can use simulations to explain it.
- Explain the way you will make inference for this model.
- How would you obtain reliable confidence intervals?
- Provide a test to check zero-inflation, if necessary use Monte Carlo tests.

3. Simulation

Simulate data for each model selecting some parameters and show that your inference algorithm is able to estimate those initial parameters.

4. Results

Show the results for the data provided for both applications.

5. Conclusion and discussion

Present:

- the main conclusions of your work,
- advantages and disadvantages,
- discuss the results obtained for you application,
- discuss future work.

References

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):265–285.

Breslow, N. E. and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88(421).

McCulloch, C. E. (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 92(437):162–170.